The application of molecular dynamics simulation techniques and free energy calculations to predict protein-protein and protein-ligand interactions

Pieffet, Gilles

Chapter 1

Introduction

This chapter provides a general introduction to proteins, protein folding and the techniques, both experimental and computational, commonly used for their study. The protein folding problem is also posed. As all the simulations in this thesis are performed using molecular dynamics simulation techniques, a special emphasis is placed on this method. The concept of free energy is discussed within the framework of protein folding. The methods used to calculate relative free energies are also introduced.

Proteins are at the center of most, if not all, biological processes. Their range of activity spans from receptor activation and signal transduction to regulating cellular processes such as membrane fusion. One class of proteins called enzymes can be seen as molecular factories catalyzing very specific chemical reactions. Proteins can also be found in membranes where they regulate the transport of specific molecules such as ions and metabolites across cellular boundaries. A common feature in all these cases is that proteins carry out their function through binding and interaction with specific partners.

Our view of proteins continues to evolve over time. At the beginning of the last century, proteins were generally thought to be simple colloids. The first crystal structure of myoglobin solved in 1958 [1] revealed that proteins could have a complex well defined 3D structure. The fact that all of the initial proteins studied in detail had structures solved by x-ray crystallography lead to the dogma that all functional proteins had well defined structures. However, numerous proteins have recently been found to lack intrinsic structure under physiological conditions [2]. It appears that they only become structured upon binding to a target molecule. Among other advantages, such a mechanism would confer the ability to bind, maybe in different conformations, to several different targets. This poses the question of the nature of the native structure. In early studies [3] it was already questioned whether the native structure of a protein (its functional form) corresponded to the thermodynamic equilibrium structure or whether the functional form was just transient and existed only for the period of time needed for the protein to perform its specific function.

In most cases, however, it does appear that the native structure of a protein is unique and directly related to its function. For this reason much effort has gone into understanding the nature of the protein structure. The structure adopted by a protein is the result of a complex molecular recognition mechanism that depends on the cooperative action of many weak non-bonding interactions (van der Waals and Coulombic). Since all the information necessary for a protein to fold to a unique structure is solely contained in the amino acid sequence [3], many methods have been developed which attempt to predict the folding behavior of a protein on the basis of its sequence only. No simple solution has yet been found to this extremely complex problem. In addition, the specific structure observed under certain conditions (pH, presence of salt...) depends not only on the sequence but also on the nature of the environment increasing the complexity of structure prediction.

1.1 Proteins

Proteins are biological macromolecules. They consist of a chain of amino acids (or residues) linked by peptide bonds. There are 20 naturally occurring amino acids (Table 1.1). The residues are composed of two parts, a backbone and a side-chain. The backbone is identical for all residues with the exception of proline.

Table 1.1: *The 20 amino acids found in nature together with their three and one letter code.*

Residue			Residue		
Alanine	Ala	A	Leucine	Leu	L
Arginine	Arg	R	Lysine	Lys	K
Asparagine	Asn	N	Methionine	Met	M
Aspartate	Asp	D	Phenylalanine	Phe	F
Cysteine	Cys	C	Proline	Pro	P
Glutamine	Gln	Q	Serine	Ser	S
Glutamate	Glu	E	Threonine	Thr	T
Glycine	Gly	G	Tryptophane	Trp	W
Histidine	His	H	Tyrosine	Tyr	Y
Isoleucine	Ile	I	Valine	Val	V

All the differences observed in the structures of different proteins are therefore determined by the side-chains.

Protein structure can be represented at different levels. The primary structure corresponds to the sequence of amino acids. The secondary structure refers to the formation of local structure [4] and describes the conformation of the backbone. Characteristic elements are α-helices and β-sheets and the structure of a protein is often described in terms of these two elements of secondary structure. It should be noted that structures described by secondary structure elements are local in space and primarily involve sequential residues (α-helices) or complementary sequences (β-sheets). Tertiary structure refers to the 3D structure or the spatial arrangement of the elements of secondary structure.

1.2 The protein folding problem

As stated above, the key to understanding the function of a protein is to be able to determine its structure. The protein folding problem can generally be defined as knowing the relationship between the amino-acid sequence and the native structure. The problem may be more clearly understood if expressed in terms of two separate aspects. The first concerns the determination of the native structure of the protein from the amino-acid sequence only. The second is related to understanding the folding process itself. As will be seen later in this chapter, a method capable of correctly predicting the final structure does not necessarily yield any insight into the mechanism of folding itself. In the same way that knowing the fold/conformation of a protein does not allow one to predict the

amino-acid sequence.

1.3 Experimental structure determination

The most important methods used to determine the structure of a protein are x-ray crystallography and NMR spectroscopy [5] and most of the structures available from public databases such as the protein data bank were obtained using these methods. Each of them has strengths and weaknesses. X-ray crystallography yields high resolution but (almost) no dynamical information. NMR spectroscopy in contrast usually yields less precise structures (the time resolution of the method is such that the signal corresponds to a time and ensemble average of structures) but offers the possibility to extract some information on the dynamical properties of the system.

1.4 Molecular dynamics

Molecular dynamics is the method of choice when one wants to study the dynamical properties of a system in full atomic detail, provided that the properties are observable within the time scale accessible to simulations. Time scale is one of the two main limitations of the method as will be discussed later. Molecular dynamics simulations are also useful when the system cannot be studied by the experimental methods mentioned above. For example when the protein cannot be crystallized or is too big or insoluble to be studied by NMR.

To calculate the dynamics of the system, that is the position of each atom as a function of time, Newton's classical equation of motion are solved iteratively for each atom:

$$F_i = m_i a_i = m_i \frac{d^2 r_i}{dt^2} \tag{1.1}$$

The force on each atom is the negative of the derivative of the potential energy with respect to the position of the atom:

$$F_i = -\frac{\partial V}{\partial r_i} \tag{1.2}$$

If the potential energy of the system is known then, given the coordinates of a starting structure and a set of velocities, the force acting on each atom can be calculated and a new set of coordinates generated, from which new forces can be calculated. Repetition of the procedure will generate a trajectory corresponding to the evolution of the system in time.

The accuracy of the simulations is directly related to the potential energy function used to describe the interactions between particles. In molecular dynamics, a classical potential energy function is used that is defined as a function of

the coordinates of each of the atoms. The potential energy function is separated into terms representing covalent interactions and non-covalent interactions. The covalent interactions may be described by the following terms:

$$V_{bond} = \sum_{i=1}^{N_b} \frac{1}{2} k_i^b (r_i - r_{0,i})^2 \tag{1.3}$$

$$V_{angle} = \sum_{i=1}^{N_\theta} \frac{1}{2} k_i^\theta (\theta_i - \theta_{0,i})^2 \tag{1.4}$$

$$V_{dihedral} = \sum_{i=1}^{N_\phi} \frac{1}{2} k_i^\phi \cos(n_i(\phi_i - \phi_{0,i})) \tag{1.5}$$

$$V_{improper} = \sum_{i=1}^{N_\xi} \frac{1}{2} k_i^\xi (\xi_i - \xi_{0,i})^2 \tag{1.6}$$

which correspond to two, three, four and four body interactions, respectively. These interactions are represented by harmonic potentials for the bond lengths r_i, for the bond angle θ_i, and for the improper dihedral (out of the plane) angle ξ_i and by a more complex potential for the dihedral angles ϕ_i. The non-covalent interactions, which correspond to interactions between particles separated by more than three covalent bonds are usually described by Coulomb's law

$$V_{Coulomb} = \sum_{i<j} \frac{1}{4\pi\varepsilon_0\varepsilon_r} \frac{q_i q_j}{r_{ij}} \tag{1.7}$$

for the electrostatic interactions and by a Lennard-Jones potential

$$V_{LJ} = \sum_{i<j} \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \tag{1.8}$$

for the Van der Waals interactions where r_{ij} is the atomic distance between particle i and j. The complete set of parameters used in the potentials (force constants, ideal bond lengths, bond angles, improper dihedral angles, dihedral angles, partial charges and Van der Waals parameters) to describe the interactions between different particle types is called the force field.

Molecular dynamics is a very useful tool. It can provide a wealth of detailed information on the structure and dynamics of proteins and peptides. However, it suffers certain limitations. First, the method is computationally very demanding and depending on the size of the system simulation times are currently limited to hundreds of nanoseconds or a few microseconds at most. For example the cumulated simulation time of the peptide studied in the second chapter, the EPO

mimetic peptide 1, is approximately 2.4 μs. This is too short to observe, for instance, the complete folding of a protein which occurs on a time scale ranging from milliseconds to seconds. Also, the form of the potential energy function must be kept simple for reasons of efficiency. The possibility to observe certain properties is directly related to the quality of the force field and, whether or not it has been parameterized for the system simulated. The quality of the force field is especially critical in the simulation of proteins. Proteins are in general only marginally stable. The difference in free energy between the folded and unfolded form is in the order of 10-20 k_bT which corresponds to the energy associated with the formation of a couple of hydrogen bonds in vacuum. The force field thus needs to be very accurate to discriminate between different conformations. However, it is questionable whether an empirical force-field can achieve the required accuracy especially when important effects such as polarization of the atoms by their environment is not taken into account by the electrostatic potential. The last limitation is that a classical description of the particles is used. This prohibits the study of quantum-mechanical based phenomena such as electron transfer or bond breaking/formation. It should be noted that mixed QM/MM mixed methods exist that can treat this type of phenomena but due to the computational cost of including a quantum treatment for part of the system, the simulation times are restricted to hundreds of picoseconds. Such simulations are restricted to essential QM process as, for example, the study of electron transfer.

Before finishing this section on MD, a few other theoretical methods that can be used to study proteins should be mentioned. The Monte Carlo (MC) method [6, 7] was historically used before MD. Monte Carlo procedures also involve the evaluation of a potential energy but differ in that an ensemble of conformations is generated by performing random displacements of the atomic positions from one conformation to the other, accepting or rejecting these based on the Metropolis criteria. Its main advantage is that it allows crossing (hopping over might be a more accurate image) of high-energy barriers provided that they are narrow. The method is also very efficient in sampling low or medium density systems but not dense systems such as proteins in solution. The main disadvantage with respect to MD is that the dynamics of the system is lost and no insight can be gained for instance on folding pathways. Homology modeling can reliably predict the fold of a protein if its sequence is close enough (>25% identity) to the sequence of a protein with a known structure. However, even when the method is successful in predicting the correct fold it still does not give any information on the nature of the interactions, on the pathways or dynamics leading to the folding of the protein and therefore does not provide any insights on the physics involved in the folding process.

1.5 Free energy

The free energy is a thermodynamic function that determines the equilibrium of a system. It is related to many if not all the physical properties a chemist or a biochemist might find of interest such as binding constants or conformational preferences. The free energy is in a way the key to the folding problem as in most cases it is believed that the native state of a protein corresponds to its lowest free energy state. A very popular view of the general folding mechanism is that there is an overall bias in the free energy towards the native state which is represented by a funnel in the free energy landscape when plotting the configurational entropy and the configurational energy as a function of a progress variable (for example the number of native contacts) [8, 9, 10]. This would explain why a protein would only visit a fraction of the conformational space before folding into its native state, solving Levinthal's paradox [11].

However it has been found using a simple model that a small penalty term applied to locally incorrect bond configurations can reduce dramatically the conformational space really accessible to proteins [12]. In this model, the protein was described by the states of N bonds connecting N+1 amino acids, each bond being characterized by only two states, correct (c) or incorrect (i). Changes in the system were made through to the conversion of a bond from c to i with a rate k_0 or from i to c with a different rate k_1. Applying a small energy penalty for making an incorrect bond, it was found that the lowest (free) energy conformation (when all the bonds are correct) could be obtained within a biologically relevant time scale. Using another simple model, it was also shown that the free energy landscape does not necessarily need to have a funnel like shape or other properties that have been proposed by some to be relevant [13] such as a large energy gap between the native and the lowest non native structure.

The free energy is usually expressed as the Helmholtz free energy, F, for an isothermal-isochoric system (the corresponding ensemble is referred to as the canonical ensemble) or the Gibbs free energy, G, for an isothermal-isobaric system respectively.

Using statistical mechanics, the Helmholtz free energy can be expressed in terms of the canonical partition function Z:

$$F = -k_B T \ln Z \tag{1.9}$$

where Z is defined as

$$Z = \frac{1}{h^{3N} N!} \int \int e^{-H(p,r)/k_B T} dp\, dr \tag{1.10}$$

for a system of N indistinguishable particles. The 3N-dimensional vectors r and p respectively correspond to the coordinates and conjugate momenta of all the particles of the system. Each pair (r, p) represents one point in the phase space of the system defined by all possible values of r and p. It can be seen that from the

definition of the partition function Z the absolute free energy can usually not be calculated as it requires the sampling of the complete phase space of the system. What can be determined is the free energy difference between two states of a system.

The relative free energy between two states A and B of a system is given by:

$$\triangle F_{BA} = F(B) - F(A) = -k_B T \ln \frac{Z_B}{Z_A} \tag{1.11}$$

which corresponds to the probability of finding the system in one state with respect to the other. Calculating the free energy with this method can be extremely inefficient depending on the type of process studied. In the case of the binding of two molecules, many association/dissociation events must be sampled in order to obtain reliable statistics on the process. Unfortunately, for strongly interacting systems, the rate of dissociation can be too slow to be simulated. This will be discussed in detail in relation to the dimerization of the EPO mimetic peptide 1 studied in the next chapter. However, the method can be successfully used to study the folding-unfolding thermodynamics of small peptides in rapid equilibrium for which conformational preferences can be calculated [14].

The free energy difference between two states A and B of a system can also be calculated as the work done on the system to force the transition from one state to the other. Standard methods to calculate free energy are the Thermodynamic Integration (TI) method [15] and the free energy perturbation (FEP) method [16]. Both make use of the so-called coupling parameter approach where the state of the system is coupled to a parameter λ. More precisely, the Hamiltonian is defined as a function of this coupling parameter λ which connects both the initial and end states such that $\mathcal{H}(\lambda_A) = \mathcal{H}_A$ corresponds to state A and $\mathcal{H}(\lambda_B) = \mathcal{H}_B$ to state B.

If the Hamiltonian is made a function of λ, the free energy also becomes a function of λ. In this case, the relative free energy between the two states A and B can be expressed as:

$$\triangle F_{BA} = F(\lambda_B) - F(\lambda_A) = \int_{\lambda_A}^{\lambda_B} \frac{\partial F(\lambda)}{\partial \lambda} d\lambda \tag{1.12}$$

$$= \int_{\lambda_A}^{\lambda_B} \left\langle \frac{\partial \mathcal{H}(\lambda)}{\partial \lambda} \right\rangle_\lambda d\lambda \tag{1.13}$$

where $\langle ... \rangle_\lambda$ represents an average over the ensemble at the corresponding λ value. Formula 1.13 is referred to as the thermodynamic integration formula [15]. TI calculations can be performed according two different schemes. The integration can be performed continuously while slowly changing the coupling parameter λ from λ_A to λ_B during the course of the simulation (slow growth method). This scheme is usually not used as the system lags behind the changing Hamiltonian and never equilibrates appropriately [17]. The other scheme is to perform simulations at

certain λ points and to evaluate the integral numerically. This way the convergence of the simulations can be checked independently at each λ point and extra λ points can be added if needed. This method is used in chapter 3 to evaluate the relative affinity of several ligands to two different serine proteases and in chapter 4 to evaluate the relative stability of a swapped dimer (SUC1) upon mutation. The method is very demanding as equilibrium simulations must be performed at intermediate states during which a representative ensemble must be sampled [18]. It should be noted that a complete sampling of the conformational space is not needed, even though results are still directly related to the extent of the phase space sampled.

Combining equation 1.11 with the coupling parameter approach leads to:

$$\triangle F_{BA} = F(\lambda_B) - F(\lambda_A) = -k_B T \ln \frac{Z(\lambda_B)}{Z(\lambda_A)} \tag{1.14}$$

$$= -k_B T \ln \left\langle e^{-[\mathcal{H}(\lambda_B)-\mathcal{H}(\lambda_A)]/k_B T} \right\rangle_{\lambda_A} \tag{1.15}$$

which has the form of an ensemble average over state A. Formula 1.15 is known as the (free energy) perturbation formula [16]. Although the method is formally exact (there is no assumption made as to the size of the perturbation and, as opposed to perturbation theories in statistical mechanics, there is no truncated expansion), it theoretically requires the sampling of the complete phase space of the reference state and therefore poses the same problem as the calculation of absolute free energy. In practice, however, convergence can be obtained if low energy conformations can be sampled both for the reference and the perturbed state which means that conformations sampled in the reference state A also have a high probability in the perturbed state B. In order to have significant overlap of the low energy regions of both ensembles, the perturbation must be small. For this reason the change between A and B is usually expressed as a sum over a series of small changes from λ to $\lambda + \Delta\lambda$:

$$\triangle F_{BA} = -k_B T \sum_{\lambda=\lambda_A}^{\lambda_B} \ln \left\langle e^{-(\mathcal{H}_{\lambda+\Delta\lambda}-\mathcal{H}_\lambda)/k_B T} \right\rangle_\lambda \tag{1.16}$$

which is usually referred to as the multi windows free energy perturbation method. Alternatively, in an effort to increase the efficiency of the method (make it less computationally demanding), another approach was derived where the sampling of the reference state was biased by the use of a soft core interaction site at positions where atoms were to be created or deleted [19]. This results in the extension of the phase space sampled in the reference state to relevant parts of the configuration space accessible to the system in the perturbed state. The increase of the conformational overlap between the two states leading to the convergence of the ensemble average. Using this soft core potential, accurate estimates of the

relative free energy can be obtained from the single ensemble of a reference state (single step perturbation) [19, 20, 21, 22].

1.6 Outline

Discovering the structure of a protein is usually an important step towards the understanding of its function and of the mechanism by which the function is carried out. However, as already stated previously, knowing the structure of a protein does not give any insight on the way the protein folds. Further, its interactions with other proteins in order to carry out its function can only be extrapolated unless the structure of the different proteins in complex together is also available. An alternative is to gain a better understanding of the interactions between proteins and between proteins and ligands in order to be able to predict how proteins or some of their elements associate with one another. This thesis discusses how these interactions lead to the formation of secondary structure elements. It also addresses the prediction of the relative affinity between proteins and between proteins and ligands using free energy calculations.

The ultimate goal when studying proteins is of course the resolution of the protein folding problem. Due to the complexity of the problem however, this has not been done yet and probably no single thesis would be able to bring a global solution to this problem. Instead, the subject of this thesis is the study of proteins and protein-protein interactions. A particular focus in the different chapters is put on the time scale needed to obtain statistically reliable information and properties.

Chapter 2. The EPO mimetic peptide 1 is used as a model system to help understand how β-sheets can rearrange by observing how the dimer can switch between different dimeric states. The interactions are the same as those involved in protein folding and the study provides insight into how secondary structure elements can find and recognize each other in the course of the folding of a protein. The study also provides insight into the time scale on which these processes happen.

Chapter 3. The importance of convergence in the case of free energy calculations using TI is assessed. It is shown that the simulation of specific intermediate points can require up to 20 ns to reach convergence in the case of simple ligands mutated in a water environment. The mutations performed on these ligands involve the simultaneous creation/deletion of many sites and illustrate both the power and limitations of these types of calculations.

Chapter 4. Reliability and applicability of TI free energy calculations when applied to protein-protein interactions are evaluated by determining the relative stability of the dimer of the Suc1 protein upon mutation. The mutations are performed on a swapped dimer and on the corresponding monomer in

order to evaluate their effect on the relative stability of the dimers. Comparison with experiment gives insight into the current state of the method and the progress that still remains to be accomplished before it can be used in a practical manner to predict protein self-assembly.

Chapter 5. Factors that affect the convergence properties of free energy calculations are investigated. The sources of errors related to the convergence of the calculations (the sampling error and the statistical error) are evaluated using three mutations of the Suc1 protein. The ability to reliably determine the error associated to the calculations is critical to the meaning and applicability of the calculations hence the two most popular methods used to calculate the statistical error are reviewed and their reliability assessed.

Chapter 6. Conclusion and outlook of this thesis.